

발음기반 번역기

Pronunciation-based Translator

딥러닝 2조

Tee Tran
teetran.com

Try Pitch



TOPIC

INPUT: "와따시와 아나따오 아이시떼이마스"



PROCESS: "私はあなたを愛しています"



OUTPUT: "나는 당신을 사랑합니다"

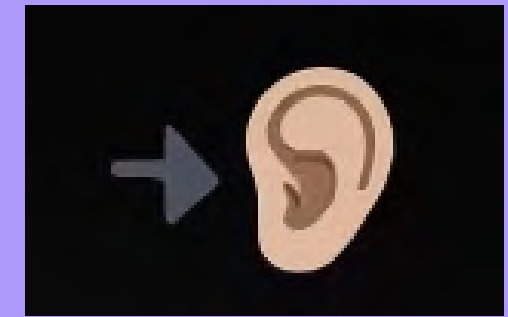
WHY???



CHAT_GPT CANNOT SOLVE

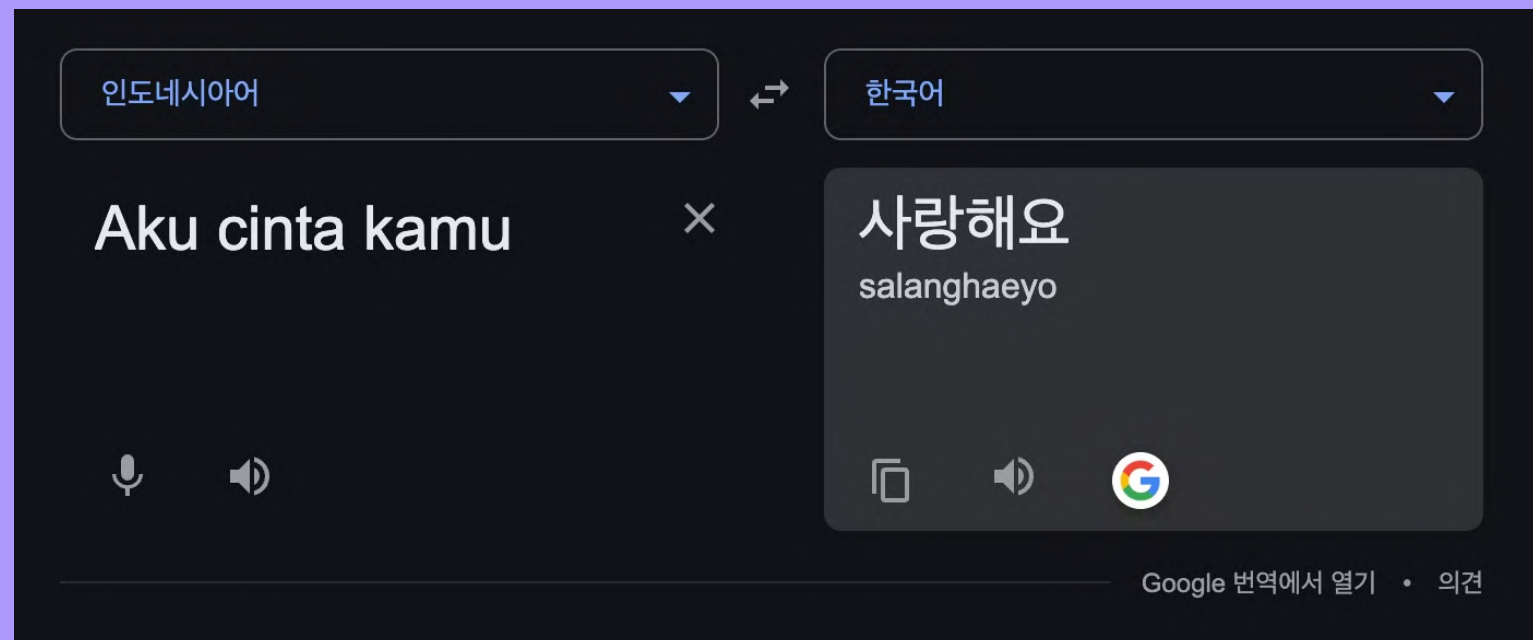
WHY???

"내가 들은건 이거였는데 왜 이런 뜻이 나와???"



result of typing as *heard*

linguistic speculation



아꾸 친타 카무

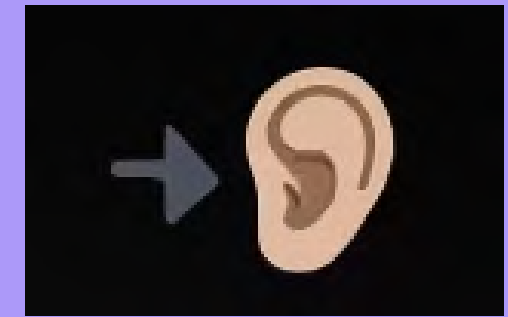
좋아, 이번에 말한 ***"아꾸 친타 카무"***는 리듬이나 발음상 일본어일 가능성이 높아 보여!
조금 더 들어보면, 아래와 같은 표현이 연상돼:

🔍 발음 추정:

- "あくちんたかむ" → 이걸 그대로는 일본어 단어가 아니지만,
"あく" (aku, 열다/나쁘다)
"ちんた" (chinta, 느릿느릿 / 한가함 - 속어)
"かむ" (kamu, 물다)

이렇게 나뉘지기도 해. 다만 자연스러운 문장은 아니기 때문에 창작 표현이거나,
아니면 다른 문장의 변형일 수 있어.

WHY???



Teaching which country's language it is

아꾸 친타 카무

좋아, 이번에 말한 *****아꾸 친타 카무*****는 리듬이나 발음상 **일본어**일 가능성이 높아 보여!
조금 더 들어보면, 아래와 같은 표현이 연상돼:

🔍 발음 추정:

- "あくちんたかむ" → 이걸 그대로는 일본어 단어가 아니지만,
"あく" (**aku**, 열다/나쁘다)
"ちんた" (**chinta**, 느릿느릿 / 한가함 – 속어)
"かむ" (**kamu**, 물다)

이렇게 나뉘지기도 해. 다만 자연스러운 문장은 아니기 때문에 **창작 표현**이거나,
아니면 다른 문장의 변형일 수 있어.

오! *****아꾸 친타 카무*****가 **인도네시아어**였다면, 이제 완전 명확해졌어 😊

🔍 발음 기반 → 원문 복원:

- 입력: 아꾸 친타 카무
- 복원: **Aku cinta kamu**
- 해석: "나는 너를 사랑해"

인도네시아어야

PRIOR STUDY

01

Machine Transliteration (Knight & Graehl, 1998)

Pronunciation-based character mapping is required for conversions such as **Romaji → Hiragana**, **Romaji → Hangul**, or **Hangul ↔ Hiragana**.

02

**Umoren et al. (2015)
"Grapheme-To Phoneme, Phoneme-To-Grapheme (G2P-P2G) Transcription Machine for Tone Languages: A Database Approach"**

Unlike standardized phonetic symbols (such as SAMPA), which do not allow arbitrary input, this project adopts an arbitrary Korean pronunciation input method.

03

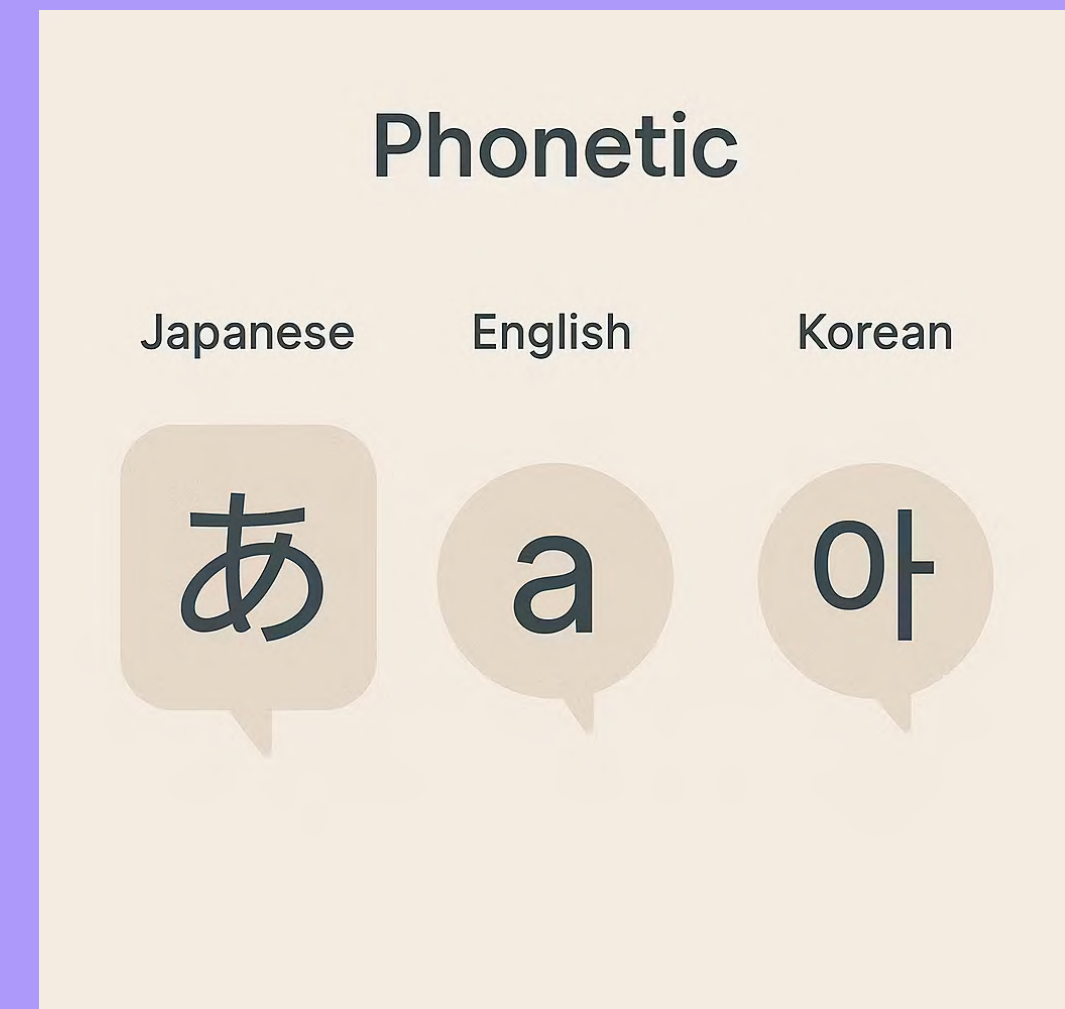
A Survey of Grapheme-to-Phoneme Conversion Methods (Appl. Sci. 2024)

(Pronunciation → Converting to Characters).

WHY JAPANESE???



1. In the case of English, typing is relatively convenient
2. Japanese is difficult to type on digital devices



3. Japanese includes a complex writing system based on Chinese characters, it is fundamentally composed of phonetic scripts such as Hiragana and Katakana. Like English and Korean, it has a structure where words are written based on how they sound.

System Architecture

1

Korean Pronunciation → Romanized text

아리가토 고자이마스 → aligato gojaimaseu

korean-romanizer

2

Romanized text → Hiragana

aligato gojaimaseu → ありがとうございます

seq2seq

3

Hiragana → Korean meaning

ありがとうございます → 고맙습니다

mbart50 or m2m100

Dataset

1

Collect Japanese sentences

→ Collect sentences from [internet](#).

2

Hiragana conversion

→ Translate Japanese sentences into **Hiragana** (ex: MeCab, fugashi, etc. use morpheme analyzer)

3

Hiragana → Roman text

→ Hiragana romanization.

ex)

m o o n a k u n a	もう泣くな。
-------------------	--------

Romanization

g2pk module

```
>>> from g2pk import G2p
>>> g2p = G2p()
>>> g2p("어제는 날씨가 맑았는데, 오늘은 흐리다.")
어제는 날씨가 맑았는데, 오느른 흐리다.
```

- applicate phonological rule

hangul_romanization

```
>>> from hangul_romanize import Transliter
>>> from hangul_romanize.rule import academic
>>>
>>> transliter = Transliter(academic)
>>> print(transliter.translit(u'안녕하세요'))
annyeonghase-yo
```

- Convert korean Text to Roman text

Current progress

1

Korean Pronunciation → Romanized text

아리가토 고자이마스 → aligato gojaimaseu

korean-rominizer

2

Romanized text → Hiragana

aligato gojaimaseu → ありがとうございます

rule-based for test

3

Hiragana → Korean meaning

ありがとうございます → 고맙습니다

mbart-large-50-many-to-many

test result

#1

들리는 대로 입력하세요: *곰엔나사이*
발음법칙 적용: 고멘나사이
로마자 변환: **gomennasai**
히라가나 변환: ごめんなさい
한국어 번역 결과: 죄송합니다.

#2

들리는 대로 입력하세요: *와따시와 아나따오 아이시떼마스*
발음법칙 적용: 와따시와 아나따오 아이시떼마스
로마자 변환: **wattasiwa anattao aisittemaseu**
히라가나 변환: わとたすいわあなとたおあいすいとてませう
한국어 번역 결과: "나와와" "아와 와"

Future Plans

- 1 Pronunciation-based Data Augmentation**

- 2 seq2seq training for Romaji → Hiragana**

- 3 Fine-tuning mBART for Hiragana → Korean meaning**

Planning for Long-term Scalability

→ Expansion to multiple languages beside Japanese including Spanish, English, and others.



Q&A



Want to make a presentation like this one?

Start with a fully customizable template, create a beautiful deck in minutes, then easily share it with anyone.

Create a presentation (It's free)